Published in final edited form as:

Biometrics. 2013 December ; 69(4): . doi:10.1111/biom.12070.

Testing for the multivariate stochastic order among ordered experimental groups with application to dose–response studies

Ori Davidov

Department of Statistics, University of Haifa, Mount Carmel, Haifa 31905 Israel,
davidov@stat.haifa.ac.il

Shyamal Peddada

Biostatistics Branch, National Institute of Environmental Health Sciences, Alexander Drive, RTP,
NC 27709, peddada@niehs.nih.gov

Abstract

The comparison of two or more ordered experimental groups based on multivariate data is common in a variety of applications such as toxicology, clinical trials and drug development, to name just a few. In this paper we develop a nonparametric methodology for analyzing such data. In particular we propose a global K sample nonparametric test for order among vector valued outcomes. The testing procedure can also be used in a post–hoc fashion to answer questions about the ordering of subgroups and/or single outcomes within any subset of experimental groups. Such a methodology does not currently exist. In contrast with standard methodology such as multivariate analysis of variance (MANOVA), and its nonparametric analogues, we do not assume that the groups differ only by a location parameter or that the components of the response vector have the same marginal distributions between and across groups, i.e., we allow for the shape of the distribution to change across groups. We emphasize that our test compares the outcome distributions, not just their mean tendencies, and explicitly incorporates and exploits the order constraints. Consequently it is more powerful than the existing unordered tests. The methodology is illustrated using genotoxicity data where the effect of hydrogen peroxide exposure on damage to DNA is evaluated using a comet assay.

Keywords

dose–response studies; nonparametric tests; ordered experimental conditions; order restricted statistical inference; stochastic order relations

1. Introduction

In many applications, researchers are interested in comparing two or more naturally ordered experimental groups. In clinical trials and in toxicology researchers often conduct dose–response studies and/or time course experiments to make comparisons between two or more experimental groups in terms of several correlated outcomes. For example, in genotoxicity studies researchers are interested in studying the extent of DNA damage caused by varying doses of a toxin over time. This is often done using comet assays. Comet assays are conducted by embedding cells, previously exposed to a test substance, in agarose on a microscopic slide. The cells are lysed leaving only a nucleoid of DNA. The nucleoids are then subject to electrophoresis. Nucleoids with intact DNA maintain their spherical shape, while if there is DNA damage, such as breaks in the double strand, the DNA leaks from the nucleoid producing a comet shaped image having an elliptical head and elongated tail. Summary measures for describing the geometry of the image and hence the extent of the leak and thus the damage to DNA are collected. These include the percent tail DNA, the

extent tail moment, the olive tail moment and the tail extent (cf. Dunson et al. 2003). The problem of interest is to compare the dose groups in terms of these correlated outcomes.

Although the outcome variables are often dependent researchers typically use univariate methods for analyzing the data (e.g., Dunson et al. 2003, NTP 2003). In doing so, there is not only a potential loss of power but also inflated Type I error (cf. Johnson and Wichern, 1998). For this reason multivariate methods are preferable. In many applications not all outcome variables are normally distributed. Also, as often seen in toxicological data, the shape of the distribution may vary by dose (Dunson and Peddada, 2008). We illustrate this point by considering the comet assay data which are discussed in greater detail in Section 5. In Figure 1 we plot the estimated densities of four outcome variables. In each panel we plot the densities of the control and four dose groups. The plots suggest that the data are not normally distributed. Furthermore, the distribution seems to vary across dose groups. Consequently standard multivariate methods, such as MANOVA, which are designed for testing the equality of central tendency may lack power. Moreover in many experimental settings the experimental groups are ordered in some sense, e.g., a toxicologist may be interested in the toxicity of a chemical at different doses or different durations of exposure, so the usual multivariate nonparametric methods (cf., Oja 2010) may not be appropriate.

The ordering of the groups induces a set of constraints on the parameter space. It is well known that tests for order are more powerful than their unordered counterparts, cf. Silvapulle and Sen (2005) as well as Sasabuchi et al. (2003) and Sasabuchi (2007). Although there is considerable literature on testing for stochastic order among univariate distributions the literature dealing with inference on nonparametric multivariate ordered distributions is sparse. Sampson and Whitaker (1989) and Lucas and Wright (1991) developed estimation and testing procedures focusing on bivariate ordinal distributions. The major obstacle to the development of an appropriate methodology has been that the nonparametric comparison of two multivariate random vectors requires comparison of their distributions over all upper sets in $\mathbb{R}^p$. The number of upper sets, however, increases super-exponentially with p (Davidov and Peddada, 2011). Hence it appears that testing for the multivariate stochastic order may not be practical even for modest p . One way to deal with this challenging problem is to analyze each variable separately and then apply an appropriate multiple testing correction. Such methods, however, do not take advantage of the underlying dependence structure. Furthermore, apart from procedures such as the Bonferroni procedure (which can be very conservative), most multiple testing procedures control the error rate by assuming that the underlying test statistics satisfy conditions which may not be easy to verify. Davidov and Peddada (2011) took the first step in providing a framework for analyzing multivariate ordered data but their work was limited to multivariate binary data. The present paper provides a solution to the general problem of multivariate ordered inference. It uses a different approach for: (i) quantifying the strength of an ordering; and (ii) dealing with the multivariate order restrictions. The resulting methodology incorporates the underlying dependence structure and reduces a very high dimensional, computationally challenging problem to one that can be addressed with available tools without compromising on the true hypotheses of interest. The proposed methodology can also be viewed as a multivariate, order restricted, extension of the tests of Jonckheere (1954) and Terpstra (1952) and of Terpstra and Magel (2003).

We start, in Section 2, by developing a methodology for comparing $K = 2$ multivariate populations. We show that the proposed methodology is easy to understand and use regardless of the dimension of the data and that it has two important properties, namely, consistency and monotonicity of power. In Section 3 the proposed methodology is extended to comparing $K > 2$ multivariate populations. The primary goal of this paper is a global test of order among K populations, however, analogous to classical multivariate analysis (e.g.

Johnson and Wichern 1998), we propose methods for post-hoc testing between pairs of populations and selected variables. The performance of the proposed procedure is numerically evaluated in Section 4, and in Section 5 we illustrate the methodology by applying it to genotoxicity data obtained from Dr. Jack Taylor (Epidemiology Branch, NIEHS). We conclude the paper in Section 6 with a brief summary and a discussion of some interesting open problems in this area. All proofs and additional theoretical results are provided in the Supplementary Materials.

2. Testing for order: the two sample problem

Let $\mathbf{X}$ and $\mathbf{Y}$ be $p \times 1$ response vectors (RVs) associated with low and high dose groups, respectively. The RV $\mathbf{X}$ is said to be smaller than $\mathbf{Y}$ in multivariate stochastic order, denoted $\mathbf{X} \preceq_{st} \mathbf{Y}$, provided that $\mathbb{P}(\mathbf{X} \in U) \leq \mathbb{P}(\mathbf{Y} \in U)$ for all upper sets $U \in \mathbb{R}^p$ (Shaked and Shanthikumar 2007). Recall that a set U is called an upper set if $\mathbf{u} \in U$ implies that $\mathbf{v} \in U$ whenever $\mathbf{v} \leq \mathbf{u}$ component-wise. If for some upper sets the inequality is strict we say that $\mathbf{X}$ is strictly smaller than $\mathbf{Y}$ in the multivariate stochastic order, i.e., $\mathbf{X} \prec_{st} \mathbf{Y}$.

2.1. Formulation

Assume first that $\mathbf{X} \preceq_{st} \mathbf{Y}$. The objective is to test the hypotheses:

$$H_0: \mathbf{X} =_{st} \mathbf{Y} \text{ versus } H_1: \mathbf{X} \prec_{st} \mathbf{Y}, \quad (2.1)$$

which specifies that the distributions of $\mathbf{X}$ and $\mathbf{Y}$ are equal under the null or strictly ordered under the alternative. Recently, Davidov and Peddada (2011) demonstrated that if $\mathbf{X} \preceq_{st} \mathbf{Y}$ then $\mathbf{X} \prec_{st} \mathbf{Y}$ if and only if $X_j \prec_{st} Y_j$ for some $1 \leq j \leq p$. This statement holds true for all multivariate random vectors. As a consequence we can invoke the union–intersection principle by rewriting (2.1) as

$$H_0: \bigcap_{j=1}^p [X_j =_{st} Y_j] \text{ versus } H_1: \bigcup_{j=1}^p [X_j \prec_{st} Y_j]. \quad (2.2)$$

This formulation is useful because it shows that the dimensionality of the problem can be reduced substantially. In this paper we exploit this result to develop tests for continuous distributions. To do so first define $\mathbf{P} = (P_1, \dots, P_p)^T$ where for $j = 1, \dots, p$ we set $P_j = \mathbb{P}(X_j \leq Y_j)$. Note that under the null $X_j =_{st} Y_j$ so $P_j = 1/2$ and if $X_j \prec_{st} Y_j$ then $P_j > 1/2$. Moreover if $X_j \preceq_{st} Y_j$ and $P_j > 1/2$ then $X_j \prec_{st} Y_j$. We emphasize that P_j is a scale independent measure of the strength of the ordering between X_j and Y_j . Thus we may reformulate (2.2) as

$$H_0: \mathbf{P} \in \mathcal{P}_0 \text{ versus } H_1: \mathbf{P} \in \mathcal{P}_1 \setminus \mathcal{P}_0 \quad (2.3)$$

where $\mathcal{P}_0 = \{\mathbf{P}_0\}$ and $\mathbf{P}_0 = (1/2, \dots, 1/2)$ is a singleton and $\mathcal{P}_1 = \{\mathbf{P} : \mathbf{P} \geq \mathbf{P}_0\}$ is a pointed convex polyhedral cone, often called the orthant cone. Hence in our formulation the ordering is captured by the vector $\mathbf{P}$. If $\mathbf{P}$ is “large” we conclude that the distributions are strongly ordered. Note that without the assumption that $\mathbf{X} \preceq_{st} \mathbf{Y}$ the hypotheses (2.3) describes a multivariate Pitman or precedence order (Arcones et al. 2002). This type of ordering may be of interest on its own. It is well known (e.g., Peddada, 1985, Khattree and Peddada, 1987) that the univariate Pitman (or precedence) order is weaker than the ordering of moments hence a test based on it may be more powerful than a comparison of means.

We emphasize that our approach reduces, reformulates, and simplifies a very high dimensional nonparametric problem of comparing two DFs over all upper sets in $\mathbb{R}^p$ to an essentially parametric testing problem as expressed in (2.3).

2.2. The proposed methodology

Let $X_1, \dots, X_{n_1}$ and $Y_1, \dots, Y_{n_2}$ be random samples from the two populations. Define

$\hat{\mathbf{P}} = (\hat{P}_1, \dots, \hat{P}_p)^T$ where for $j = 1, \dots, p$

$$\hat{P}_j = \frac{1}{n_1 n_2} \sum_{s=1}^{n_1} \sum_{t=1}^{n_2} \mathbb{I}_{(X_{js} \leq Y_{jt})} \quad (2.4)$$

is the nonparametric maximum likelihood estimator of $P_j = \mathbb{P}(X_j \leq Y_j)$. Note that (2.4) is the well known Wilcoxon–Mann–Whitney (WMW) statistic. Using multivariate U–Statistic theory (cf. Chapter 5 of Kowalski and Tu, 2008) it can be established that if $n_1, n_2 \rightarrow \infty$ and

$n_1/(n_1 + n_2) \rightarrow \lambda \in (0, 1)$ then $\sqrt{n}(\hat{\mathbf{P}} - \mathbf{P}) \Rightarrow N_p(0, \Sigma)$ where $n = n_1 + n_2$ is the total sample size and $\Rightarrow$ denotes convergence in distribution. It can be further shown that the

elements of Σ denoted σ_{ij} , are given by $\sigma_{ij} = \lim_{n \rightarrow \infty} n \text{Cov}(\hat{P}_i, \hat{P}_j)$ which equals

$(\Gamma_{1,ij} - P_i P_j) / \lambda + (\Gamma_{2,ij} - P_i P_j) / (1 - \lambda)$ where $\Gamma_{1,ij} = \mathbb{P}(X_i \leq Y_i, X_j \leq Y_j)$ and

$\Gamma_{2,ij} = \mathbb{P}(X_i \leq Y_i, X'_j \leq Y'_j)$ here $\mathbf{X}, \mathbf{X}'$ are IID with distribution function (DF) F and $\mathbf{Y}, \mathbf{Y}'$ are IID with DF G . Conditions for the positivity of all σ_{ij} are given in the Supplementary

Materials. Note that $\Gamma_{1,ij}$ may be estimated by the $(n_1 n_2 (n_2 - 1))^{-1} \sum_{s,t \neq v} \mathbb{I}_{(X_{is} \leq Y_{it}, X_{js} \leq Y_{jv})}$ and similarly for $\Gamma_{2,ij}$ whereas P_i and P_j are estimated by (2.4). Therefore each σ_{ij} and consequently Σ can be consistently estimated from the data. However, since σ_{ij} are individually estimated the resulting estimator for Σ may be an indefinite matrix. To avoid such problems, we apply the bootstrap methodology for estimating the variance matrix of $\hat{\mathbf{P}}$.

The bootstrap estimator of $\hat{\Sigma}$ has several advantages over the plug-in estimator. It is guaranteed to be positive definite and surprisingly, it may in fact be more accurate in small samples because it accounts for variance terms of order $O(n^{-1})$ neglected by the large sample approximations (cf., Singh 1981). Finally, but very importantly, the bootstrap estimator can be easily generalized to the multi-sample setting described in Section 3.

Motivated by the common strategy used in constrained inference problems (cf. Silvapulle and Sen, 2005), we propose testing (2.3) using the statistic D_n given by

$$D_n = n \left\{ \left\| \hat{\mathbf{P}} - \Pi_{\hat{\Sigma}_0}(\hat{\mathbf{P}} | \mathcal{P}_0) \right\|_{\hat{\Sigma}_0}^2 - \left\| \hat{\mathbf{P}} - \Pi_{\hat{\Sigma}_0}(\hat{\mathbf{P}} | \mathcal{P}_1) \right\|_{\hat{\Sigma}_0}^2 \right\}. \quad (2.5)$$

where $\|\mathbf{x}\|_V \equiv (\mathbf{x}^T V^{-1} \mathbf{x})^{1/2}$, $\hat{\Sigma}_0$ estimates the scaled variance of $\hat{\mathbf{P}}$ under the null and $\bar{\mathbf{P}} = \Pi_{\hat{\Sigma}_0}(\hat{\mathbf{P}} | \mathcal{P}_1) = \arg \min \left\{ (\hat{\mathbf{P}} - \mathbf{P})^T \hat{\Sigma}_0^{-1} (\hat{\mathbf{P}} - \mathbf{P}) : \mathbf{P} \in \mathcal{P}_1 \right\}$ is the projection of the unconstrained estimator $\hat{\mathbf{P}}$ onto the convex cone $\mathcal{P}_1$ with respect to the matrix $\hat{\Sigma}_0$. The value of $\bar{\mathbf{P}}$ is readily obtained by quadratic programming. Observe that (2.5) simplifies to $n \left\{ (\hat{\mathbf{P}} - \mathbf{P}_0)^T \hat{\Sigma}_0^{-1} (\hat{\mathbf{P}} - \mathbf{P}_0) - (\hat{\mathbf{P}} - \bar{\mathbf{P}})^T \hat{\Sigma}_0^{-1} (\hat{\mathbf{P}} - \bar{\mathbf{P}}) \right\}$. Note that (2.5) has the form of a likelihood ratio statistic (LRT) for an asymptotically normal random vector (Silvapulle and Sen, 2005). The null distribution of the proposed statistic is given below.

Theorem 2.1. If $\mathbf{P} = \mathbf{P}_0$ and $n_1, n_2 \rightarrow \infty$ with $n_1/(n_1 + n_2) \rightarrow \lambda \in (0, 1)$ then

$$\lim_{n \rightarrow \infty} \mathbb{P}(D_{\mathbf{n}} \geq x) = \sum_{k=1}^p w_k(\boldsymbol{\Sigma}_0, \mathbf{I}) \mathbb{P}(\chi_k^2 \geq x) \quad (2.6)$$

for some weights $w_k(\boldsymbol{\Sigma}_0, \mathbf{I})$, $k = 1, \dots, p$ discussed below.

Thus (2.5) follows a chi-bar distribution in large samples. The weights $w_k(\boldsymbol{\Sigma}_0, \mathbf{I})$, $k = 1, \dots, p$ are non negative and sum to unity; they depend on the variance under the null and the restriction on $\mathbf{P}$ imposed by the alternative. In general the weights can not be computed in closed form if $p > 4$. However they can be estimated by Monte-Carlo methods. Moreover the entire distribution may also be approximately simulated to obtain critical values. In practice however it is often more convenient to test (2.1) using the bootstrap or permutation distribution of $D_{\mathbf{n}}$. Finally, if $p = 1$ the statistics $D_{\mathbf{n}}$ reduces to the one sided WMW test.

2.3. A union–intersection test

Formulating (2.1) as (2.2) suggests a union–intersection test (UIT) as an alternative to (2.5). UITs for a variety of univariate problems in order restricted inference are discussed by Sen (2007). In general projection tests such as $D_{\mathbf{n}}$ are expected to have high power. However, as the dimension of the problem increases and if the correlation among the components of $\hat{\mathbf{p}}$ is high then the estimated variance matrix is likely to become unstable. This may potentially impact the power of $D_{\mathbf{n}}$. In these situations, the UIT test may serve as a better alternative.

Define for $j = 1, \dots, p$ the statistic $T_{\mathbf{n},j} = \sqrt{n}(\hat{P}_j - 1/2)$ where $\hat{P}_j$ is defined in (2.4).

Clearly $T_{\mathbf{n},j}$ can be used to test the hypothesis $H_0^{(j)}: X_j = Y_j$ versus $H_1^{(j)}: X_j \prec_{st} Y_j$. Thus the UIT is

$$T_{\mathbf{n}} = \max_{1 \leq j \leq p} T_{\mathbf{n},j} \quad (2.7)$$

and the null is rejected if $T_{\mathbf{n}} > c_{\alpha}$ where the critical value solves $\mathbb{P}(T_{\mathbf{n}} > c_{\alpha}) = \alpha$. Clearly $T_{\mathbf{n}} \Rightarrow \max_{1 \leq j \leq p} U_j$ where $(U_1, \dots, U_p)$ is distributed as $N_p(0, \boldsymbol{\Sigma}_0)$. Thus, for small p , critical values can be computed exactly, otherwise they can be easily computed by simulation. The test (2.7) is based on combining p WMW tests. The WMW is an attractive choice because of its interpretability and good power properties against a variety of alternatives. However it is worth mentioning that each $T_{\mathbf{n},j}$ could be replaced with a different rank test. It is well known that rank tests are well suited for testing for univariate stochastic order (Hajek et al. 1999, Davidov 2012). More generally though any test for univariate stochastic order can be used. The consistency and monotonicity of the proposed tests is discussed in the Supplementary Materials.

3. A multi–sample extension

As stated in the introduction, this paper is motivated by the problem of comparing K ordered experimental conditions in terms of p dependent response variables. The underlying premise is that $\mathbf{X}_1 \preceq_{st} \dots \preceq_{st} \mathbf{X}_K$ where $\mathbf{X}_i$, $i = 1, \dots, K$ are p dimensional RVs and the problem of interest is to test following hypothesis:

$$H_0: \mathbf{X}_1 =_{st} \dots =_{st} \mathbf{X}_K \text{ versus } H_1: \mathbf{X}_1 \leq_{st} \dots \leq_{st} \mathbf{X}_K \quad (3.1)$$

with at least one strict inequality under the alternative.

Although (3.1) is a natural generalization of (2.1) from $K = 2$ to $K > 2$ samples there is no unique way of generalizing the test statistic. Initially we considered generalizing (2.4) using p , K -sample (univariate) statistic of the form,

$$\hat{P}'_j = \left(\sum_{i < i'} n_i n_{i'} \right)^{-1} \sum_{i < i'} \sum_{s=1}^{n_i} \sum_{t=1}^{n_{i'}} \mathbb{I} \left(X_{ijs} \leq X_{i'jt} \right), \quad j=1, \dots, p. \quad (3.2)$$

where X_{ijs} is the s^{th} observation ($s = 1, \dots, n_i$) on the j^{th} component ($j = 1, \dots, p$) in the i^{th} group, $i = 1, \dots, K$. This statistic was proposed by Jonckheere (1954) and Terpstra (1952) and is referred to as the JT test. The JT test is known to have good power for most configurations in the alternative. However like many other test procedures in order restricted inference it is not very specific when neither the null nor the alternative are true. For example, if the true ordering is $X_{Kj} \prec X_{1j} \prec_{st} \dots \prec_{st} X_{K-1j}$ then the JT test may potentially reject the null in favor of the alternative even though this configuration is not part of the alternative. The problem arises because the union of the null and the alternative is not the entire parameter space.

Specifically, the violation of a monotonic dose response relationships at low doses is known in the toxicological literature as hormesis (Calabrese and Baldwin, 2002). Simulation results provided in Section 4, suggest that the test based on (3.2) has a high probability of rejecting the null hypothesis in the presence of hormesis. This property of the test based on (3.2) is undesirable for the application at hand. To deal with this problem we employ the statistic

$$\hat{P}_j = \left(\prod_{i=1}^K n_i \right)^{-1} \sum_{i_1=1}^{n_1} \dots \sum_{i_K=1}^{n_K} \mathbb{I} \left(X_{1ji_1} \leq X_{2ji_2} \dots \leq X_{Kji_K} \right) \quad (3.3)$$

introduced by Terpestra and Magel (2003) who considered a K sample univariate problem. Their test, referred to as the TM test, insists on simultaneously using all the inequalities specified by the alternative. They demonstrated that their test is substantially more specific than JT, i.e., it rejects the null less often than JT when the true configuration is neither the null nor the alternative hypothesis, while not losing much power when the alternative is true. Consequently the test based on (3.3) deals better than the tests based on (3.2) with the problem of hormesis and thus is more useful in this application. Note that $\hat{P}_j$ estimates $P_j = \mathbb{P} \left(X_{1j} \leq X_{2j} \leq \dots \leq X_{Kj} \right)$ for $j = 1, \dots, p$ and $\mathbf{P} = (P_1, \dots, P_p)^T$ as in Section 2.1.

Proposition 3.1. If $X_1, \dots, X_K$ are independent univariate RVs and $X_1 \prec_{st} \dots \prec_{st} X_K$ then $\mathbb{P} \left(X_1 \leq X_2 \leq \dots \leq X_K \right) \geq 1/K!$ with equality if and only if $X_1 =_{st} \dots =_{st} X_K$.

Proposition 3.1 implies that the hypotheses (3.1) may be rewritten as (2.3) where $\mathcal{P} = \{\mathbf{P}_0\}$ and $\mathbf{P}_0 = (1/K!, \dots, 1/K!)$ is again a singleton and $\mathcal{P}_1 = \{\mathbf{P} : \mathbf{P} \geq \mathbf{P}_0\}$ is a pointed convex polyhedral cone. We proceed as in Section 2. By appealing to U-Statistic theory it can be

shown that $\sqrt{n} \left(\hat{\mathbf{P}} - \mathbf{P} \right) \Rightarrow N_p(0, \Sigma)$ where $n = \sum_i n_i$, $n_i \rightarrow \infty$ and $n_i/n \rightarrow \lambda_i > 0$. A straightforward, albeit tedious, calculation shows that the matrix Σ has entries

$$\sigma_{ij} = \sum_{k=1}^K (\Gamma_{k,ij} - P_i P_j) / \lambda_k \text{ where}$$

$$\Gamma_{k,ij} = \mathbb{P} \left(X_{1i} \leq \dots \leq X_{ki} \leq \dots \leq X_{Kj} \leq \dots \leq X_{Kj} \right) \text{ and } X_{si} \text{ is}$$

independent of X'_{sj} whenever $s \neq k$. These expression generalize those obtained in Section 2.2.

It follows that (3.1) can be tested by employing the statistic (2.5) properly modified to the K sample setting. We shall estimate the variance by bootstrap methodology to circumvent the challenges related to the estimation of the above covariance matrix, especially when the sample sizes are not large. Analogues of the UIT (2.7) are also readily available. Consistency and monotonicity of these tests may be argued as before.

In applications where hormesis, or more generally the specificity of the test, is not a substantive issue one could employ a multivariate test based on (3.2) instead of (3.3). This option may be attractive when K is large. In such situation the probabilities P_j , described in Proposition 3.1, may be very small (recall that their null value is $1/K!$) and it may not be desirable to conduct inference on such small probabilities. In contrast the statistic based on (3.2) has expected value $1/2$ for all K and thus may be more appropriate for large K . We denote the corresponding statistics based on (3.2) by D'_n and T'_n respectively.

3.1. Post-hoc tests

Once the global null hypothesis is rejected researchers may be interested in performing post-hoc analysis to determine which pairs of experimental conditions are different and which outcomes, or groups of outcomes, are ordered by the level of the exposure. In addition for a given pair of experimental conditions, researchers may be interested in performing tests on individual outcomes. Thus a variety of post-hoc analyses can be performed. Numerous post-hoc methods have been proposed in the literature in the context of normal theory multivariate analysis of variance (MANOVA). For example, if the researcher does not decide in advance which specific linear hypothesis to be test, then one may typically use critical regions based on Scheffe's simultaneous confidence ellipsoid. On the other hand, if the research is interested in some prechosen linear combinations then one may use Bonferroni's rectangular critical regions (cf. Johnson and Wichern 1998). In the present context of nonparametric testing of multivariate distributions, one may adopt a similar strategy. As noted in Davidov and Peddada (2013), the generalization of Scheffe's method (or Roy's largest root test) to nonparametric multivariate problems is nontrivial. However, Bonferroni's method can be directly applied in the present context by modifying the critical values suitably.

4. Simulations

We evaluated the Type I error and power of D_n and T_n (as well as D'_n and T'_n) for testing hypotheses (3.1) on p -variate normal populations where for $i = 1, \dots, 4$, where $X_i \sim N_p(\mu_i, \Sigma)$. A random sample of size $n_i = 20$ was selected from each population and we considered two values of p , ($p = 3$ and 6). We assumed that $\Sigma = (\rho^{|i-j|})$ has a serial correlation matrix with four values for the serial correlation coefficient ρ ($0, 0.25, 0.5, 0.75$). We define $0^0 = 1$ so $\rho = 0$ corresponds to the independence case. Without any loss of generality, the Type I errors for the test statistics were estimated by taking $\mu_i = 0$ for $i = 1, \dots, 4$. Powers of the tests were estimated for different choices of μ_i , $i = 1, \dots, 4$. The mean patterns used for power calculations are displayed in Table 1. All Type I errors and powers were estimated using 1000 simulation runs and the sample covariance matrices were estimated using 200 bootstrap samples. In Table 2 Type I errors and powers for D_n and T_n as well as D'_n and T'_n are reported. The Type I errors are generally close to the nominal level 0.05, although in some instances they tend to be conservative. As expected, D_n tends to be more powerful than T_n for smaller values of ρ . For larger values of ρ , T_n tends to perform better than D_n . A similar phenomena is observed for D'_n and T'_n . Note that in almost every case the power of D_n and T_n is lower than the power of D'_n and T'_n respectively. This is the price paid by D_n and T_n for correcting for hormesis. However, for large values of ρ the tests D_n and D'_n have

comparable power. It is also interesting to note that the power of D'_n and T'_n decreases more rapidly with ρ compared with D_n and T_n .

Table 3 reports on the probability of rejecting the null for the mean patterns $M_5 - M_8$ reported in Table 1. These patterns represent cases where the alternative hypothesis is not true. As expected in every case, the probability of rejecting the null using the JT based test is substantially higher than the with the TM based test. Thus, as expected, the TM based tests are more specific than the JT based tests. In fact, it can be easily argued that as the low dose mean goes to negative infinity the probability of rejecting the null hypothesis goes to zero by our test, whereas the rejection probability for the JT based test would go to unity. Thus, as desired, the proposed test is very specific while the JT based test is not.

5. An Illustration

To determine the genotoxicity of hydrogen peroxide (H_2O_2), Dr. Taylor's lab (Epidemiology branch, NIEHS) conducted a dose response study using the comet assay and measured several variables, namely the percent tail DNA (V_1), the extent tail moment (V_2), the olive tail moment (V_3) and the tail extent (V_4) immediately after exposure, 60 minutes after exposure and 90 minutes after exposure. These times are denoted by t_0 , t_{60} and t_{90} respectively. The doses of H_2O_2 used in the study were in micromoles (denoted μM) 0, 5, 20, 50, and 100. These doses are denoted by d_0 , d_5 , d_{20} , d_{50} and d_{100} . There were 100 cells in each dose groups except for the 50 μM group where data on only 50 cells were available.

For simplicity of illustration, in Figure 1 we provided the empirical marginal densities of each of the four outcomes for the 5 μM dose 60 minutes after exposure. As we see from Figure 1, none of the variables appears to be normally distributed in any of the dose groups. We performed Anderson–Darling tests for normality on each variable in each dose group. In almost every case the p-value was < 0.005 . Thus, marginal normality is rejected overwhelmingly in all dose groups for all variables. For each variable the mean, median and standard deviation along with p-value based on Anderson–Darling test for normality are provided in Table S1 (see the Supplementary Materials).

The theoretical expectation is that cells will recover over time from exposure to hydrogen peroxide. This means that damage to DNA is expected to be high at t_0 and decrease over time. Hence there is a scientific interest in testing, at each dose level $d \in \{d_0, d_5, d_{20}, d_{50}, d_{100}\}$, the hypotheses:

$$H_0: V_{d,t_0} =_{st} V_{d,t_{60}} =_{st} V_{d,t_{90}} \text{ versus } H_1: V_{d,t_0} \succ_{st} V_{d,t_{60}} \succ_{st} V_{d,t_{90}} \quad (5.1)$$

with at least one strict inequality under the alternative. Here $V_{d,t} = (V_{d,t,1}, V_{d,t,2}, V_{d,t,3}, V_{d,t,4})$ where $V_{d,t,j}$ is the RV corresponding to the j^{th} outcome variable at the t^{th} time point and d^{th} dose. Results using the D_n test are summarized in Table 4. For a visual description of the results, corresponding to each significant dose of hydrogen peroxide, we plotted in Figure 2 the median values at each time point for the four response variables relative to their median at t_0 . The ratios are multiplied by 100 to obtain the percent change relative to time t_0 . As expected, there was no significant trend in the control group overtime, i.e., the null hypothesis in (5.1) was not rejected for the control group. However, the null in (5.1) was overwhelmingly rejected in all other dose groups with highly significant p-values (< 0.0001) (Table 4). For those dose groups where the multivariate null in (5.1) was rejected, we performed post-hoc analysis for each variable. Thus, within each dose group (excluding the control) we tested the following univariate hypotheses: $H_{0j}: V_{d,t_0,j} =_{st} V_{d,t_{60},j} =_{st} V_{d,t_{90},j}$ versus $H_{1j}: V_{d,t_0,j} \prec_{st} V_{d,t_{60},j} \prec_{st} V_{d,t_{90},j}$ where $j = 1, 2, 3, 4$. To adjust for multiple testing within each dose, we used Bonferroni corrections so that the overall family wise error rate

was controlled at 5%. In every dose group, we found a significant trend (p -value < 0.0001) over time in all variables except for the tail extent (i.e., V_4) (see Table 4). Our results are consistent with biologist's belief that DNA repair mechanisms are able to reverse damage to DNA (at least for the dose levels considered in this study).

Biologists are also interested in testing whether there is a dose-response at each time point. Theory predicts that initially there will be a strong dose-response relationship but that over time, due to the DNA repair mechanisms, this relationship will weaken and finally disappear. Thus we tested hypotheses analogous to (5.1) but flipping the role of dose and time. Results of this analyses are presented in Table S2 and a graphical summary in Figure S1. As expected, there was a highly significant multivariate dose response at time point t_0 (p -value < 0.0001), a moderately significant response at time t_{60} (p -value = 0.0456) and no significant dose-response at t_{90} (p -value 0.788).

Since there is an almost an instantaneous response to hydrogen peroxide (at time t_0), we performed the step-down procedure described by Davidov and Herman (2012), applied here to random vectors, with the goal of identifying the smallest dose which yields a significant response at t_0 . At the first step test $H_0 : V_{d_0, t_0} =_{st} \cdots =_{st} V_{d_{100}, t_0}$ versus $H_1 : V_{d_0, t_0} \prec_{st} \cdots \prec_{st} V_{d_{100}, t_0}$. If the null is rejected then we drop the highest dose group and test $H_0 : V_{d_0, t_0} =_{st} \cdots =_{st} V_{d_{50}, t_0}$ versus $H_1 : V_{d_0, t_0} \prec_{st} \cdots \prec_{st} V_{d_{50}, t_0}$. If the null is rejected then we continue by dropping V_{d_{50}, t_0} and so forth. The last dose d , where $d > d_0$, at which we reject the null is declared to be the minimal toxic dose. Note that each hypotheses is tested at the 5% level of significance. As discussed in Davidov and Herman (2012), and the references therein, this step-down procedure controls the overall Type I error at 5% level. Using this procedure we find that there is a significant DNA damage at the lowest dose considered in the study, namely, 5 μ M. This is true for both the multivariate outcome and also for each of the variables V_1 , V_2 and V_3 separately.

Finally we note that Dunson and Peddada (2008) performed a univariate Bayesian analysis using only the olive tail moment (V_3) as the response variable. They concluded that at t_0 there was a highly significant increase in DNA damage between 0, 5, 20 μ M dose groups but further increase in DNA damage at higher doses was unclear. We reanalyzed the data using our methods and discovered that although there was no significant increase in DNA damage from 50 μ M to 100 μ M dose groups, moreover we discovered a highly significant increasing trend in DNA damage from 20 to 50 to 100 μ M. The significantly increasing trend was observed not only in the multivariate response vector (V_1 , V_2 , V_3 , V_4) (p -value < 0.0001) but also in the individual variables V_1 , V_2 , V_3 . For each of these variables, the p -values after Bonferroni adjustment was < 0.0001 .

6. Concluding remarks

In many applications researchers are interested in comparing two or more ordered multivariate populations. The ordering may be with respect to the stage of a tumor, the dose of a chemical, the duration of an exposure and so forth. In such experiments the responses may not be multivariate normally distributed. In fact, as mentioned herein the shape of the distribution may change with dose. In such cases, the standard MANOVA based methodology may not be appropriate. A natural framework for ordering multivariate distributions is to compare populations in terms of the multivariate stochastic order. However, as noted in Davidov and Peddada (2011), testing for multivariate stochastic order is challenging.

In this paper we considered a strategy which exploits the fact that a researcher conducting a dose-response study intrinsically assumes that the response is monotone in the dose. This

assumption substantially reduces the computational burden while preserving the underlying structure in the data and essentially reduces a highly nonparametric problem to a parametric one. A nice aspect of our multi-sample extension is that the dimension of the problem does not grow with the number of groups compared. Our comparison of the marginal distributions is based on the WMW test which is a rank based test which does not compare the means but rather compares distribution functions over their entire support. It is also important to note that in the absence of any a priori assumptions on the ordering of the data, our procedure tests for a multivariate Pitman or precedence order (Arcones et al. 2002). An alternative approach was also recently proposed by Davidov and Peddada (2013) who studied directed inference for multivariate ordered distributions but, in contrast with this paper, focused on the linear stochastic order.

In some situations, such as in toxicological experiments, researchers discover a “U” (or an “inverted U”) shaped dose response curve where the up-turn (or down-turn) may occur in the low dose. Toxicologists often refer to this phenomena as hormesis (Calabrese and Baldwin, 2002). In such data, standard trend tests may erroneously declare a significant monotonic trend. Paradoxical situations such as these arise commonly in the field of order restricted inference. The main problem is that often the union of the null and the alternative hypotheses is not the entire p dimensional Euclidean space. To deal with this problem Terpestra and Magel (2003) proposed a clever approach to the problem of testing for monotonic trends by using a test statistic that carefully ignores portions of the sample space where the monotonic constraint is violated. In this paper we adopted their strategy for testing multivariate stochastic order. Thus, our proposed methodology is less likely to declare a “U” (or inverted “U”) shaped response to be a monotonic response, i.e., it is robust with respect to hormesis.

Although several interesting questions remain to be explored, this paper takes the first step towards a practically useful formulation of the problem of testing for multivariate stochastic order. There are several open problems that remain to be addressed in this area. In this paper we assumed the underlying random variables to be continuous. However, in many instances researchers obtain a combination of continuous and discrete responses. Our methodology readily extends to analyze such data. For example one simple approach for tied data is to

estimate P_j in (2.4) by, $\sum_{s=1}^{n_1} \sum_{t=1}^{n_2} \mathbb{I}_{(X_{js} \leq Y_{jt}, X_{js} \neq Y_{jt})} / \sum_{s=1}^{n_1} \sum_{t=1}^{n_2} \mathbb{I}_{(X_{js} \neq Y_{jt})}$, that is discarding those pairs for which $X_{js} = Y_{jt}$ for any s and t . Note that the above quantity estimates $\mathbb{P}(X_j \leq Y_j | X_j \neq Y_j)$ which under the null hypothesis equals 1/2. Thus inference can be developed along the lines outlined above. Additional methods dealing with ties are also available (cf. Conover 1973).

Supplementary Materials

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The research of Ori Davidov was partially supported by the Israeli Science Foundation Grant No. 875/09. This research [in part] was supported by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences (Z01 ES101744-04).

References

- [1]. Arcones MA, Kvam PH, Samaniego FJ. Nonparametric estimation of a distribution subject to a stochastic precedence constraint. *Journal of the American Statistical Association*. 2002; 97:170–182.

- [2]. Cai B, Dunson DB. Bayesian multivariate isotonic regression splines: application to carcinogenicity studies. *Journal of the American Statistical Association*. 2007; 102:1158–1171.
- [3]. Calabrese EJ, Baldwin LA. Defining hormesis. *Hum Exp Toxicol*. 2002; 21:91–97. [PubMed: 12102503]
- [4]. Conover WJ. On methods of handling ties in the Wilcoxon signed–rank test. *Journal of the American Statistical Association*. 1973; 68:985–988.
- [5]. Davidov O, Peddada SD. Order restricted inference for multivariate binary data with application to toxicology. *Journal of the American Statistical Association*. 2011; 106:1394–1404. [PubMed: 22973069]
- [6]. Davidov O. Ordered inference, rank statistics and combining p-values: a new perspective. *Statistical Methodology*. 2012; 9:456–465.
- [7]. Davidov O, Herman A. Ordinal dominance curve based inference for stochastically ordered distributions. *Journal of the Royal Statistical Society, Series B*. 2012; 74:825–847.
- [8]. Davidov O, Peddada SD. The linear stochastic order and directed inference for multivariate ordered distributions. *The Annals of Statistics*. 2013; 41:1–40.
- [9]. Dunson DB, Watson M, Taylor JA. Bayesian latent variable models for median regression on multiple outcomes. *Biometrics*. 2003; 59:296–304. [PubMed: 12926714]
- [10]. Dunson DB, Peddada SD. Bayesian nonparametric inference on stochastic ordering. *Biometrika*. 2008; 95:859–874.
- [11]. Hajek, J.; Sidak, Z.; Sen, PK. *Theory of Rank Tests*. Academic Press; 1999.
- [12]. Joe, H. *Multivariate Models and Dependence Concepts*. Chapman and Hall; 1997.
- [13]. Johnson, R.; Wichern, D. *Applied Multivariate Statistical Analysis*. Prentice Hall; 1998.
- [14]. Jonckheere AR. A distribution free k -sample test against ordered alternatives. *Biometrika*. 1954; 41:133–145.
- [15]. Khattree R, Peddada SD. A Short Note on Pitman Nearness for Elliptically Symmetric Estimators. *Journal of Statistical Planning and Inference*. 1987; 16:257–260.
- [16]. Kowalski, J.; Tu, XM. *Modern Applied U-Statistics*. Wiley; 2008.
- [17]. Lucas LA, Wright FT. Testing for and against stochastic ordering between multivariate multinomial populations. *Journal of Multivariate Analysis*. 1991; 38:167–186.
- [18]. NTP. NTP toxicology and carcinogenesis studies of citral (microencapsulated) (CAS No. 5392-40-5) in F344/N rats and B6C3F1 mice (feed studies). 2003.
- [19]. Oja, H. *Multivariate Nonparametric Methods with R*. Springer; 2010.
- [20]. Peddada SD. A short note on Pitman's measure of nearness. *American Statistician*. 1985; 39:298–299.
- [21]. Sasabuchi S, Tanaka K, Tsukamoto T. Testing homogeneity of multivariate mean vectors under order restrictions when the covariance matrices are common but unknown. *Annals of Statistics*. 2003; 31:1517–1536.
- [22]. Sasabuchi S. More powerful tests for homogeneity of multivariate normal mean vectors under an order restriction. *The Indian Journal of Statistics*. 2007; 69:700–716.
- [23]. Sampson AR, Whitaker LR. Estimation of multivariate distributions under stochastic ordering. *Journal of the American Statistical Association*. 1989; 84:541–548.
- [24]. Shaked, M.; Shanthikumar, JG. *Stochastic Orders*. Springer; 2007.
- [25]. Sen PK. Union–Intersection principle and constrained statistical inference. *Journal of Statistical Planning and Inference*. 2007; 137:3741–3752.
- [26]. Silvapulle, MJ.; Sen, PK. *Constrained Statistical Inference*. Wiley & Sons; 2005.
- [27]. Singh K. On the Asymptotic Accuracy of Efron's Bootstrap. *Annals of Statistics*. 1981; 9:1187–1195.
- [28]. Terpstra TJ. The asymptotic normality and consistency of Kandall's test against trend when ties are present in one ranking. *Indag Math*. 1952; 14:327–333.
- [29]. Terpstra JT, Magel RC. A new nonparametric test for the ordered alternative problem. *Nonparametric Statistics*. 2003; 15:289–301.

Figure 1. Estimated density plots for the four variables, namely, Percent tail DNA, Extent tail moment, Olive tail moment and Tail extent, at 5 μ M dose 60 minutes after exposure.

Figure 2. Median response at 60 and 90 minutes after exposure to various doses of hydrogen peroxide, relative to the median response immediately after exposure (expressed in percent). The response immediately after exposure is set to 100%.

Table 1

Patterns of means used for power calculations.

Mean Pattern	Group Means
M_1	(0,0,0), (.5, .5, .5), (.5, .5, .5), (.5, .5, .5)
M_2	(0, 0, 0), (.1, .1, .1), (.2, .2, .2), (.5, .5, .5)
M_3	(0, 0, 0, 0, 0, 0), (.5, .5, .5, .5, .5, .5), (.5, .5, .5, .5, .5, .5), (.5, .5, .5, .5, .5, .5)
M_4	(0, 0, 0, 0, 0, 0), (.1, .1, .1, .1, .1, .1), (.2, .2, .2, .2, .2, .2), (.5, .5, .5, .5, .5, .5)
M_5	(0, 0, 0), (-.5, -.5, -.5), (0,0, 0), (0,0, 0)
M_6	(0, 0, 0), (-1, -1, -1), (0, 0, 0), (0, 0, 0)
M_7	(0, 0, 0), (-.5, -.5, -.5), (0, 0, 0), (0, 0, 0), (0, 0, 0), (0, 0, 0)
M_8	(0, 0, 0), (-1, -1, -1), (0, 0, 0), (0, 0, 0), (0, 0, 0), (0, 0, 0)

Note that in pattern $M_5 - M_8$ the means are not ordered (these correspond to the hormesis hypothesis).

Table 2

Type I errors and power for the proposed procedures with nominal level $\alpha = .05$.

Type I errors						
Set Up			TM-based Test		JT-based Test	
p	Means	ρ	D_n	T_n	D'_n	T'_n
3	M_0	0.00	0.026	0.028	0.040	0.035
3	M_0	0.25	0.024	0.025	0.041	0.045
3	M_0	0.50	0.026	0.025	0.041	0.039
3	M_0	0.75	0.054	0.029	0.036	0.035
6	M_0	0.00	0.022	0.037	0.030	0.039
6	M_0	0.25	0.032	0.036	0.021	0.039
6	M_0	0.50	0.038	0.029	0.041	0.037
6	M_0	0.75	0.039	0.047	0.029	0.036
Powers						
3	M_1	0.00	0.381	0.321	0.615	0.501
3	M_1	0.25	0.353	0.337	0.478	0.459
3	M_1	0.50	0.305	0.307	0.370	0.423
3	M_2	0.75	0.261	0.286	0.283	0.391
3	M_2	0.00	0.468	0.389	0.689	0.558
3	M_2	0.25	0.410	0.397	0.547	0.514
3	M_2	0.50	0.365	0.365	0.422	0.475
3	M_2	0.75	0.305	0.342	0.316	0.437
6	M_3	0.00	0.636	0.469	0.816	0.598
6	M_3	0.25	0.425	0.357	0.609	0.535
6	M_3	0.50	0.370	0.358	0.460	0.502
6	M_3	0.75	0.306	0.333	0.294	0.445
6	M_4	0.00	0.653	0.469	0.889	0.660
6	M_4	0.25	0.532	0.429	0.705	0.608
6	M_4	0.50	0.451	0.408	0.524	0.564
6	M_4	0.75	0.351	0.372	0.333	0.501

The mean patterns are described in Table 1, the pattern M_0 denotes equality of all means.

Table 3

Probability of rejecting the null hypothesis for the proposed procedures under hormesis (recall that here the alternative is not true) with nominal level $\alpha = .05$.

Set Up			TM-based Test		JT-based Test	
p	Means	ρ	D_n	T_n	D'_n	T'_n
3	M_5	0.00	0.066	0.066	0.122	0.112
3	M_5	0.25	0.060	0.055	0.101	0.100
3	M_5	0.50	0.053	0.043	0.092	0.104
3	M_5	0.75	0.056	0.049	0.081	0.096
3	M_6	0.00	0.045	0.048	0.230	0.210
3	M_6	0.25	0.047	0.042	0.177	0.204
3	M_6	0.50	0.056	0.041	0.155	0.213
3	M_6	0.75	0.057	0.052	0.139	0.206
6	M_7	0.00	0.052	0.053	0.133	0.116
6	M_7	0.25	0.064	0.062	0.140	0.144
6	M_7	0.50	0.054	0.049	0.105	0.117
6	M_7	0.75	0.060	0.060	0.075	0.114
6	M_8	0.00	0.039	0.039	0.243	0.238
6	M_8	0.25	0.043	0.044	0.195	0.256
6	M_8	0.50	0.037	0.027	0.165	0.224
6	M_8	0.75	0.051	0.040	0.135	0.220

The mean patterns are described in Table 1.

Table 4

P-values for time trend at each dose level.

Individual Tests						
H ₂ O ₂ Dose (μM)	<i>n</i>	Global Test	PTD	ETM	OTM	TE
0	100	0.1878	0.0242	0.0475	0.0543	0.3609
5	100	< 0.0001	< 0.0001	< 0.0001	< 0.0001	0.1458
20	100	< 0.0001	< 0.0001	< 0.0001	< 0.0001	0.7677
50	50	< 0.0001	< 0.0001	< 0.0001	< 0.0001	0.5622
10	100	< 0.0001	< 0.0001	< 0.0001	< 0.0001	0.0656

Global as well as individual post-hoc tests are reported (PTD = percent tail DNA, ETM = extent tail moment, OTM = olive tail moment, TE = tail extent)